CHAPITRE I

LES CORPUS ETIQUETES

Etiqueter un texte, c'est une forme d'annotation dans laquelle on associe à des segments de texte, le plus souvent les « mots », une ou plusieurs étiquettes, le plus leur catégorie grammaticale voire leur lemme.

Dans la première section, nous donnons de brefs exemples de corpus étiquetés et nous définissons les types d'étiquetage rencontrés. Un premier exemple d'utilisation de corpus étiquetés (section 2) repose sur un étiquetage approfondi d'une partie seulement du corpus. Il vise à mettre en évidence de manière inductive une typologie des textes sur la base des corrélations observées entre les traits linguistiques retenus. Un second exemple (section 3) fait appel à un étiquetage complet mais fruste (la partie du discours et quelques renseignements morphologiques). Cet étiquetage permet de contraster les « parlures » qui coexistent dans le corpus étudié. Nous abordons en section 4 l'utilisation d'étiqueteurs ou de corpus étiquetés et en section 5 les enjeux théoriques des recherches rendues possibles par ce niveau d'annotation.

1. DEFINITIONS

Commençons par trois brefs exemples, qui donnent un aperçu de la diversité des étiquetages effectifs ... comme de leur manque de lisibilité et de clarté.

1.1 Exemples

1.1.1.1 Enfants

Les réponses fournies par les personnes interrogées :

> Les difficultés financières et matérielles.
>
> Je ne sais pas, les gens sont égoïstes, peut-être.

sont lemmatisées et étiquetées (cf. 3.2) de la manière suivante :

> <S01=23> le les {DETDEF} difficulté difficultés {NOMFP} financier financières {ADJFP} et et {CCOORD} matériel matérielles {ADJFP} . . {PONCT-FORTE}
>
> <S01=31> je je {PROPERS} ne ne {ADVNEG} savoir sais {VIPR1S} pas pas {ADVNEG} , , {PONCT-FAIBLE} le les {DETDEF} gens gens {NOMMP} être sont {VIPR3P} égoïste égoïstes {ADJMP} peut-être peut-être {ADV} . . {PONCT-FORTE}

Chaque réponse commence par des renseignements sur l'interviewé : son âge (en deuxième position après S01= : 1 renvoie à inférieur à 30 ans, 2 à entre 30 et 50 ans, 3 à au delà de 60 ans) et son niveau d'étude (en première position après S01= : 1 = sans, 2 = baccalauréat, 3 = études supérieures). Puis chaque mot, précédé de son lemme, est suivi de sa catégorie morphosyntaxique entre accolades (NOMMS = nom masculin singulier, par exemple).

1.1.1.2 Mitterrand1

Le fragment suivant est extrait de l'émission de TF1 *Ça nous intéresse, Monsieur le président* du 2 mars 1986 :

> [...] moi, je suis de la France - je ne dis pas : je suis la France - [...]

Il est codé de la manière suivante par D. Labbé (1990) :

```
moi,je,5
",",",",","p"
je,je,5
suis,être,11
de,de,81
la,le,7
France,France,22
-,-,p
je,je,5
ne,ne,6
dis,dire,11
pas,pas,6
:,:,p
je,je,5
suis,être,11
la,le,7
France,France,22
```

Le texte annoté est constitué d'une série de triplets comme suis,être,11 : le mot, le lemme, la catégorie, représentée par un nombre. Les trois informations sont séparées par des virgules.

1.1.1.3 Susanne

La phrase :

DAN MORGAN TOLD HIMSELF HE WOULD FORGET Ann Turner[1] :

est représentée ainsi :

N01:0010b	NP1m	DAN	Dan	[O[S[Nns:s.
N01:0010c	NP1s	MORGAN	Morgan	.Nns:s]
N01:0010d	VVDv	TOLD	tell	[Vd.Vd]
N01:0010e	PPX1m	HIMSELF	himself	[Nos:i.Nos:i]
N01:0010f	PPHS1m	HE	he	[Fn:o[Nas:s.Nas:s]
N01:0010g	VMd	WOULD	will	[Vdc.
N01:0010h	VV0v	FORGET	forget	.Vdc]
N01:0010i	NP1f	Ann	Ann	[Nns:o.
N01:0010j	NP1s	Turner	Turner	.Nns:o]Fn:o]S]
N01:0010k	YF	+.	-	.

Le texte est ici présenté sous la forme d'un tableau : à un mot du texte de départ correspond une ligne. Chaque ligne fournit une suite de champs. Ici pour la troisième ligne :

N01:0010d	VVDv	TOLD	tell	[Vd.Vd]

- une référence : le nom du fichier dont provient cet extrait (N01) et un numéro de ligne au sein de ce fichier : 0010d ;
- une indication d'édition : le tiret indique que le texte n'a pas été corrigé à cet endroit ;
- une catégorie : VVDv ;
- la forme fléchie telle qu'on la rencontre dans le corpus : told ;
- le lemme correspondant : tell ;
- la structure syntaxique dans laquelle s'insère le mot[2] : [Vd.Vd] indique que ce mot est la tête d'un groupe verbal. Le point signale l'endroit où le mot et sa catégorie doivent s'insérer. C'est l'équivalent de [Vd [VVdv told]].

1.2 L'inévitable éparpillement des étiquetages

Les exemples donnés manifestent la diversité en taille et en visée des jeux d'étiquettes et des stratégies d'étiquetage sous-jacentes. Cette

[1] Les majuscules sont dans le texte de départ.
[2] Nous reviendrons sur ce dernier champ au chapitre suivant, consacré aux corpus arborés : cette annotation syntaxique n'est généralement pas considérée comme faisant partie de l'étiquetage à proprement parler.

diversité tient à l'utilisation envisagée du corpus mais aussi à son mode d'étiquetage (manuel ou automatique) ainsi qu'à l'absence de consensus sur certains catégories ou sur leur extension.

L'expérience montre qu'un groupe d'annotateurs n'est pas forcément cohérent dans les étiquettes qu'il attribue manuellement à un texte. Il en va de même pour un même individu au fil du temps.

J. Véronis et L. Khouri soulignent (1995, p. 235) le fait que les jeux d'étiquettes ne sont généralement pas comparables, ce qui retarde l'évaluation ou la combinaison des étiqueteurs et des étiquetages. Pour reprendre Leech et ses collègues (1994, p. 51) : « il n'y a pas de 'meilleur jeu d'étiquettes', [...] dans la pratique la plupart des jeux d'étiquettes constituent plutôt des compromis entre la finesse de la description linguistique et ce qui peut être attendu, pour des raisons pratiques, d'un système automatique d'étiquetage[3]. » On peut recourir à un jeu d'étiquettes important pour pouvoir distinguer aisément certains cas d'ambiguïté, quitte à se ramener à un jeu plus restreint une fois l'étiquetage opéré[4]. Inversement, sur certains points, le jeu d'étiquettes peut en rester à des distinctions relativement grossières, parce qu'il s'avère difficile d'obtenir, sur des subdivisions plus fines, un consensus de la part des personnes définissant l'ensemble d'étiquettes à utiliser (Greenbaum, 1993) ou parce que des catégories trop fines rendraient plus long et plus hasardeux le travail de correction manuelle des résultats de l'étiquetage automatique. Greenbaum (*ibid.*, p. 18) donne l'exemple de la distinction comptable / non comptable, importante pour les noms en grammaire anglaise, mais difficile à établir avec sûreté, *a fortiori* à automatiser. Il propose alors de s'en tenir à l'opposition, aisément détectable, entre singulier et pluriel. À charge pour ceux qui entendent précisément étudier la dimension comptable / non comptable d'annoter en conséquence leur corpus !

Par ailleurs, les jeux d'étiquettes correspondent aussi sur certains points à des divergences théoriques réelles. Il en va de même de la projection des catégories, soulignent J. Véronis et L. Khouri (*ibid.,* p. 237) : « Même si l'on est d'accord sur le jeu d'étiquettes, leurs extensions (c'est-à-dire l'ensemble des formes lexicales qu'elles couvrent) peuvent être différentes. Le problème est particulièrement aigu pour les catégories fermées, déterminants, pronoms, adjectifs indéfinis, etc., où l'on rencontre de très grosses différences d'appréciation dans les catégories, et ce dans la plupart des langues. » Comme l'indiquent Greenbaum et Yibin (1994, p. 35) : « l'identité des étiquettes [entre deux jeux] peut être trompeuse, dans la mesure où l'assignation des étiquettes peut être différente. » Ils citent le cas de l'étiquette adverbe qui est conservée par ***ICE*** (International Corpus of English) pour les adverbes utilisés comme modifieurs de noms (*then* dans *the then president*) mais que l'étiqueteur CLAWS remplacerait par l'étiquette adjectif. Dans les cas

[3] J.-P. Chanod et P. Tapanainen (1995a) indiquent ainsi qu'ils ont ignoré la distinction masculin / féminin en français pour les noms et les adjectifs, dans la mesure où cette distinction suppose l'utilisation de contextes larges *(une envie de soleil diffuse*) et où finalement, pour leurs objectifs (repérage de l'accord sujet / verbe et ambiguïté nom / verbe), elle joue un rôle mineur.

[4] C'est la pratique d'E. Tzoukermann et de ses collègues (1995) avec des jeux de 253 et 67 étiquettes respectivement.

de conversion, c'est-à-dire de passage d'une catégorie à une autre sans changements dérivationnels, doit-on attribuer la catégorie de départ ou celle d'arrivée ? Comment catégoriser par exemple *parler* dans la séquence « le parler vrai » : comme un infinitif ou comme un nom ?

1.3 Une représentation canonique

Les corpus étiquetés peuvent donc se présenter sous des formats variables : verticalement (comme ***Mitterrand1*** ou ***Susanne***) ou horizontalement (***Enfants***). Dans ces trois exemples, la nature des informations doit être déduite de l'usage de divers caractères qui prennent un sens particulier : crochets, point, virgule, accolades, passages à la ligne, ainsi que de la place où les informations figurent. La catégorie constitue la troisième colonne de ***Susanne*** et de ***Mitterrand1*** et elle occupe la troisième position de chaque triplet pour ***Enfants***.

On peut figurer ainsi le décodage de l'étiquetage d'un mot annoté dans ***Mitterrand1*** :

mot	séparateur de champ	lemme	séparateur de champ	catégorie	séparateur de triplet
suis	,	être	,	11	passage à la ligne

Pour faciliter la récupération d'un champ donné et la transmission des corpus, on doit passer de ces indications positionnelles à une représentation logique, ce qui revient à isoler chaque type d'information et à lui donner un nom, soit avant cette information :

catégorie=verbe, lemme=être, forme=être

soit « autour » de cette information :

<catégorie>verbe</catégorie><lemme>être</lemme><forme>suis</forme>.

Cette dernière représentation, destinée à faciliter les échanges et réutilisations de corpus, repose sur des normes de balisage présentées au chapitre VI.

Ces conventions rendent explicite une représentation canonique de l'étiquetage. Les informations associées à un segment de texte peuvent en effet être représentées par une structure d'associations trait-valeur du type de celles utilisées par les formalismes syntaxiques contemporains[5]. Nous notons ces structures entre accolades, chaque trait étant séparé par le signe = de sa valeur à cet endroit et par une virgule du trait suivant. La ligne de ***Susanne*** donnée *supra*, abstraction faite du champ notant l'analyse syntaxique, se note alors :

[5] On se reportera à (Abeillé, 1993, p. 29-31) pour une présentation générale de ces structures et à (Ligozat, 1994, ch. 3 et ch. 5) pour un approfondissement formel.

{référence=N01:0010d, catégorie=VVDv, forme=told, lemme=tell}

et celle de Mitterrand1 :

suis,être,11

se transcrit ainsi :

{forme=suis, lemme=être, catégorie=11}

Comme les noms des traits sont fournis, on peut disposer les associations trait-valeur dans n'importe quel ordre. La version suivante de la ligne de de ***Mitterrand1*** est strictement équivalente à la précédente :

{catégorie=11,forme=suis,lemme=être}

Enfants ne fournit que la catégorie et le lemme, à côté des indications sur le diplôme et l'âge du locuteur. Ces indications pourraient être elles-mêmes ajoutées sous forme de traits attachés à chaque mot. Elles seraient alors « distribuées » au lieu d'être mises en facteur, ce qui donnerait, en format vertical :

{diplôme=baccalauréat, âge=60+, catégorie=DETDEF, forme=les, lemme=le}

{diplôme=baccalauréat, âge=60+, catégorie=NOMFP, forme=difficultés, lemme=difficulté}

De telles structures de traits sont « ouvertes » : il est toujours possible de rajouter des « dimensions » (par exemple des étiquettes sémantiques). On peut également enlever une partie des associations trait-valeur attachées à un « mot » et simplifier par là-même son étiquetage. On en verra un exemple dans la section 3.

Susanne fournit un trait référence identifiant de manière unique le mot examiné. Dans ***Mitterrand1***, il faut connaître le fichier dont provient l'occurrence. Le soin apporté par ***Susanne*** sur ce point peut paraître superflu. C'est pourtant en définitive sur cette identification univoque que repose la possibilité de vérifier les annotations portées sur un corpus ou les analyses qui en sont faites. Un autre chercheur peut se reporter exactement au bon endroit dans le texte de départ, examiner un contexte plus large, etc. C'est donc la condition *sine qua non* d'un travail collectif.

Si l'on adopte cette représentation canonique, on constate que le trait catégorie est utilisé différemment selon les cas. Par exemple, pour le mot *je*, la valeur de ce trait est 5, c'est-à-dire Pronom pour ***Mitterrand1*** et PROPERS pour ***Enfants***. Dans ce corpus, l'étiquette précise donc, de manière relativement transparente, le type de pronom dont il s'agit. On peut alors expliciter les composants d'une telle étiquette : {catégorie=pronom, type=personnel}. Il est fréquent que les étiquettes d'un corpus ne soient pas atomiques mais complexes : on doit les décomposer. C'est le cas pour ***Susanne***, où VVDv est en fait une abréviation pour : {categorie=verbe, temps=passé}. Développer ainsi les étiquettes complexes[6] facilite l'élagage

[6] Le projet européen MULTEXT de création de ressources linguistiques informatisées, monolingues et multilingues, et d'outils génériques d'annotation et d'exploitation de

ou l'enrichissement des traits attachés à un « mot ».

1.4 Types d'étiquetage

L'étiquetage peut être produit par un programme qu'on appelle un étiqueteur (*tagger*), ou bien résulter d'une annotation manuelle, ou bien provenir d'une combinaison des deux. Le traitement de gros volumes de textes rend cependant inéluctable le recours à un étiqueteur.

1.4.1 Etiquetage intégral ou partiel

Dans les exemples que nous avons fournis, chaque mot fait l'objet d'un étiquetage. On rencontre par ailleurs des textes étiquetés partiellement : les renseignements attachés à certains mots sont inexistants on incomplets. Il peut s'agir de limites purement techniques : l'étiqueteur utilisé bute sur des mots « inconnus », c'est-à-dire absents des dictionnaires qu'il utilise ou que ne résolvent pas les règles morphologiques qu'il emploie. Ou bien, face à un mot inconnu, l'étiqueteur fait des propositions moins précises que celles déclenchées par les mots répertoriés dans les dictionnaires employés.

L'étiquetage partiel peut aussi être visé en tant que tel. Un sous-ensemble des mots du texte est jugé pertinent pour la recherche envisagée, il est donc étiqueté, le reste est ignoré. Par exemple, si l'on entend étudier la répartition des marques de l'énonciation dans un corpus, on peut envisager un étiquetage limité aux mots retenus comme révélateurs sur ce point : embrayeurs, certains adverbiaux, indications temporelles et aspectuelles des verbes ...

1.4.2 Une étiquette ou plusieurs étiquettes

Un corpus étiqueté n'est pas forcément totalement « désambiguïsé », c'est-à-dire qu'un mot peut recevoir plusieurs étiquettes. Dans ***BNC***, à l'issue de l'étiquetage, demeurent un peu plus de 3 % de problèmes non résolus, d'« ambiguïtés », représentées par des étiquettes composites (*portmanteau tags*[7]), comme nom_verbe, pour l'hésitation entre nom et verbe. Pour un fragment de l'exemple de ***Mitterrand1*** fourni ci-dessus, un résultat non désambiguïsé serait :

{mot=je, lemme=je,		

corpus (Véronis et Khouri, 1995) insiste sur la nécessité de distinguer les descriptions lexicales, c'est-à-dire l'ensemble des associations trait-valeur qui caractérisent chaque forme, et les étiquettes, le passage des premières aux secondes se faisant par traduction, toute description lexicale devant correspondre à une étiquette au plus.

[7] Littéralement, des étiquettes-valises, sur le modèle de *portmanteau-word*.

catégorie=pronom}		
{mot=suis, lemme=être, catégorie=verbe}	{mot=suis, lemme=suivre, catégorie=verbe}	
{mot=la, lemme=le, catégorie=déterminant}	{mot=la, lemme=le, catégorie=pronom}	{mot=la, lemme=la, catégorie=nom}
{mot=France, lemme=France, catégorie=nom}		

où figurent les deux verbes correspondant potentiellement à la forme fléchie *suis* : *suivre* et *être*, tous deux légitimes hors contexte, et les trois étiquettes possibles pour *la*[8].

La degré d'étiquetage nécessaire à une expérience sur un corpus dépend étroitement des objectifs de la recherche envisagée. Si l'on veut se servir d'un corpus étiqueté pour extraire des suites de catégories syntaxiques, on peut tolérer un tel degré d'ambiguïté et trier *a posteriori* les résultats. Par contre, si l'on souhaite étudier un phénomène massif (comme la détermination) dans des gros corpus, on ne saurait se satisfaire d'un étiquetage qui laisse en suspens les choix (ici entre déterminant et pronom pour *le*, *la*, *les* ...).

1.4.3 Une vision large de l'étiquetage

Etiqueter un segment de texte (un mot, mais aussi un groupe de mots, une phrase, un paragraphe, etc.), c'est, de manière générale, lui associer des informations arbitrairement complexes[9]. Ces informations peuvent se situer à plusieurs niveaux de l'analyse linguistique : morphologie, syntaxe, sémantique, pragmatique , sans se limiter d'ailleurs aux aspects linguistiques (comme le trait diplôme utilisé pour ***Enfants*** ou le trait référence de ***Susanne***).

Cette vision élargie de l'étiquetage ne correspond cependant pas à l'acception la plus répandue. Quand on parle de corpus étiqueté, en particulier dans la communauté TALN, on fait référence le plus souvent à un document où chaque mot possède une étiquette morpho-syntaxique et une seule.

[8] Déterminant, pronom et nom (dans l'expression : *donner le la*).
[9] Nous avons fourni des structures de traits plates. Rien n'empêche d'employer des co-indiciations (Ligozat, 1994) assurant des partages de valeurs (on y a recours au chapitre suivant), ou encore des structures arbitrairement enchâssées qui regroupent des « paquets » de traits : DETMS est l'abréviation de {catégorie=déterminant, accord={genre=masculin, nombre=singulier}}, où le trait accord regroupe les traits de genre et nombre.

2. ÉTIQUETAGE PARTIEL ET TYPOLOGIE DE TEXTES

Le fait de disposer de textes partiellement étiquetés (un certain nombre de traits linguistiques fins sont privilégiés) permet d'entreprendre une typologie linguistique de ces textes, mais il n'est pas sûr qu'on puisse généraliser aisément les oppositions dégagées.

2.1 Circularité des démarches typologiques habituelles

La typologie des textes a suscité de nombreux travaux. Le plus souvent, ces recherches cherchent soit à caractériser les modes de production des textes (typologies situationnelles), soit à identifier les fonctions visées par les textes (typologies fonctionnelles). Les objectifs peuvent être didactiques (permettre à un apprenant d'identifier et de produire les différents types de textes de sa langue ou d'une langue étrangère) ou linguistiques, par exemple dans la lignée de la distinction histoire *versus* discours de Benveniste[10]. L'hypothèse partagée par ces différentes recherches est que chacun des types postulés se caractérise par l'association d'un certain nombre de caractéristiques linguistiques.

La démarche part souvent des types situationnels ou fonctionnels définis au départ, examine les textes qui relèvent de chacun de ces types et leur fonctionnement linguistique, et essaie de mettre en évidence certaines corrélations entre types et traits linguistiques. On ne sait toutefois pas si, en partant d'une autre typologie *a priori*, on ne rassemblerait pas sous un même chef des textes différents, ce qui aurait toutes chances de produire des agrégats de traits linguistiques distincts de ceux produits par la typologie précédente. La répartition des textes retenus sous les rubriques choisies est elle-même contestable. Il y a là une circularité d'autant plus gênante que l'existence de types textuels distincts paraît intuitivement fondée, même s'il s'avère délicat de l'étayer empiriquement.

2.2 Dégager les corrélations de traits linguistiques : D. Biber

Une autre optique consiste à faire émerger les types de textes grâce à un traitement statistique de textes étiquetés. C'est la ligne directrice des travaux de D. Biber (1988, 1989). Ce dernier examine les cooccurrences entre 67 traits linguistiques dans les 1 000 premiers mots[11] de 481 textes d'anglais contemporain écrit et oral. Ces textes proviennent de ***LOB*** et

[10] Elle oppose les énoncés reliés au moment de l'énonciation (emploi du présent, d'« embrayeurs » comme les pronoms de première et deuxième personne) : le *discours* à ceux qui effacent cet ancrage (emploi du passé simple, de la non-personne, c'est-à-dire la troisième personne) : l'*histoire*

[11] Cet échantillonnage a pour fonction de faciliter la comparaison des distributions de traits linguistiques. Cf. chapitre VII et chapitre IX.

London-Lund et relèvent de « genres » divers : articles de recherche, reportages, conversations, nouvelles radiophoniques ... Les traits étudiés ressortissent à 16 catégories distinctes comme marqueurs de temps et d'aspect, adverbes et locutions adverbiales de temps et de lieu, pronoms et pro-verbes, questions, passifs, modaux, coordination, négation... Ils sont identifiés automatiquement (en limitant au maximum la vérification manuelle)[12].

L'étiquetage mis en œuvre par Biber s'éloigne de l'étiquetage morpho-syntaxique pratiqué en général. Il est partiel et partial. Il est « inéquitable » : il s'intéresse à des fonctionnements linguistiques très spécifiques qu'il analyse en détail tandis qu'il en laisse d'autres dans l'ombre. Par exemple, il privilégie certains verbes (modaux) et certaines formes verbales (passif, présent ...), mais ne traite pas systématiquement l'ensemble des classes de verbes ni toutes les flexions verbales.

La statistique multidimensionnelle [13] est mise à contribution pour repérer les oppositions majeures entre associations de traits linguistiques. Elle rassemble les traits qui ont tendance à apparaître ensemble. Elle constitue dans le même temps les configurations de traits qui sont systématiquement évités par ces rassemblements. Cette démarche permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à ces constellations positives et négatives. Ces pôles deux à deux constituent des dimensions. Chaque texte, par son emploi des traits linguistiques étudiés, se situe en un point déterminé de l'espace à *n* dimensions déterminé par cette analyse.

La typologie construite par D. Biber à partir des résultats de l'analyse factorielle s'organise autour de cinq dimensions. La première oppose les textes qui se caractérisent par l'usage de *do* comme pro-verbe, celui de *be* comme verbe principal, le présent, les démonstratifs, les contractions du type *don't*, la première et la deuxième personne du singulier, le pronom *it* aux textes qui favorisent les noms, les mots longs, des adjectifs attributs, les prépositions. Biber appelle cette première dimension *production impliquée* versus *production informationnelle*[14]. Les autres dimensions sont nommées l'*orientation narrative*[15] versus *non narrative*, la *référence dépendante*[16] ou non de la situation d'énonciation, la *visée persuasive apparente*[17] *ou non*, *le style impersonnel*[18] *ou non*. Biber souligne que les dimensions proposées à l'issue de l'interprétation des contrastes majeurs mis en évidence par l'analyse factorielle sont en fait des prototypes, des pôles de fonctionnements textuels. Chacune des dimensions mises en évidence oppose deux pôles, mais les textes concrets se situent en des points variés des « échelles » ainsi définies.

A partir de ces cinq dimensions, en utilisant des techniques de classification automatique[19], Biber aboutit à huit types de textes, en

[12] Ces traits et leur repérage sont décrits en détail dans (Biber, 1988, p. 211-245).
[13] Cf. chapitre IX.
[14] *Involved* versus *informational production*.
[15] Caractérisée par le passé, la 3e personne, la négation synthétique, les participes présents.
[16] Manifestée par les adverbiaux, en particulier de temps et de lieu.
[17] Les traits privilégiés comprennent les infinitifs, les modaux, les subordonnées conditionnelles.
[18] Favorisant les passifs sans agent et les passifs avec *by*.
[19] Cf. chapitre IX.

fonction de leur place sur chacune de ces dimensions :

1) Interaction interpersonnelle intime (*intimate interpersonal interaction*) ;

2) Interaction informationnelle (*informational interaction*) ;

3) Exposé « scientifique » (*« scientific » exposition*) ;

4) Exposé savant (*learned exposition*) ;

5) Fiction narrative (*imaginative fiction*) ;

6) Récit (*general narrative exposition*) ;

7) Reportage situé (*situated reportage*) ;

8) Argumentation impliquée (*involved persuasion*).

Ces types ne correspondent pas forcément aux intuitions communes. C'est ainsi qu'on ne débouche pas sur un type unique *interaction* ou *dialogue*, mais deux : l'interaction à visée informationnelle et l'interaction à visée interpersonnelle. De la même manière, Biber distingue plusieurs types de textes « expositifs » et de textes narratifs.

2.3 Généralité des typologies induites

Cette démarche permet la construction inductive d'une typologie de textes, basée sur les corrélations effectives entre traits linguistiques[20]. Elle court néanmoins le risque d'aboutir à des oppositions qui, pour avoir été établies à partir de textes concrets, ne valent que pour ces textes et pour les traits choisis pour les opposer[21]. Peut-on accorder une portée plus générale aux types ainsi construits ? Biber (1995) a appliqué la même démarche, mais cette fois-ci à quatre corpus, le corpus anglais initial et trois ensembles de textes en coréen, somali[22] et nukulaelae tuvaluan[23]. Malgré des différences nettes, liées en particulier au degré d'alphabétisation et à la place des traditions orales dans les langues considérées, Biber (*ibid.*, p. 359) pense pouvoir émettre l'hypothèse que les types textuels qu'il dégage sont communs à plusieurs langues, mêmes si leur réalisation linguistique diffère d'une langue à l'autre.

L'articulation de ces constats généraux, sur des corpus diversifiés, avec des analyses dans un domaine particulier ne va cependant pas de soi. Ainsi, Bergounioux *et coll.* (1982) étudient les résolutions générales votées par les congrés confédéraux des quatre centrales interprofessionnelles, CFDT, CFTC, CGT et FO, pendant les années 1971-1976. Ce corpus n'est pas étiqueté, soulignons-le. La répartition précise d'un certain nombre de formes (marques d'énonciation, détermination, coordination, pronoms, prépositions, etc.) dans les textes

[20] J.-P. Sueur (1982) étudie dans une optique très proche les contrastes entre parties de la Résolution Générale du congrès de 1976 de la CFDT. Il étiquette (manuellement cette fois) les traits qui lui paraissent pertinents et utilise là encore l'analyse factorielle des correspondances pour mettre en évidence les oppositions majeures.

[21] Il est intéressant à cet égard de comparer les traits retenus par Biber avec ceux choisis par Sueur (1982) et ceux privilégiés par Bronckart (1985).

[22] Langue parlée par environ 5 millions de personnes en Somalie, à Djibouti, en Éthiopie et au Kenya.

[23] Langue parlée par 350 personnes sur l'atoll Nukulaelae du groupe Tuvalu (Pacifique).

de ces quatre organisations syndicales a pour objectif de dégager l'organisation d'ensemble de ces textes (*ibid.,* p. 169-186). Un programme qui isole les mots qui sont significativement sur-employés dans une partie d'un corpus au regard de leur emploi dans le corpus entier[24] est utilisé pour évaluer les phénomènes étudiés. Ce programme dégage en même temps les sous-emplois significatifs d'une partie au regard du tout[25]. Les convergences des sur-emplois et des sous-emplois permettent d'opposer (*ibid.,* p. 175) une structure dite analytique, utilisée par la CFDT et la CGT à une structure dite déclarative, préférée par FO et la CFTC. Le premier type de résolution sur-emploie en particulier le verbe *être* à la troisième personne de l'indicatif présent, les modaux, les pronoms à la première personne du pluriel et les possessifs de même personne, les pronoms de troisième personne. Le deuxième type sur-emploie les verbes déclaratifs (*appelle*, *considère*, *estime*, *exige* ...), ayant pour sujet *le congrès* ou le sigle (*la CFTC*), suivis d'une complétive en *que*. Une autre étude (Habert, 1983), consacrée aux résolutions générales des congrès de la CFTC de 1945 à 1964 et de la CFDT de 1965 à 1979[26], trouve une opposition similaire. D'un côté une résolution « circonstancielle », ancrée dans le temps de l'énonciation : indications précises de lieu, verbes d'affirmation ou d'interpellation. De l'autre une résolution « théorique » qui s'affranchit de l'*ici et maintenant* de l'énonciation : présent de vérité générale (avec les flexions d'*être* et *avoir*), effacement de l'énonciateur, verbes modaux, marques d'articulation logique du discours, etc. Les résolutions examinées se situent entre ces deux pôles, la résolution « théorique » prenant le pas en 1945, moment d'affirmation du syndicalisme chrétien dans une France de l'après-guerre marquée par le rôle du Parti Communiste et de la CGT, et en 1970, 1973 et 1976 où la CFDT, après 1968, opte pour le *socialisme autogestionnaire*[27]. À travers ces deux études, l'une sur une période courte (5 ans), l'autre sur le moyen terme (34 ans), il semble que deux types de textes, au moins, soient disponibles pour permettre à un acteur social de se situer dans le présent, associés à des « postures » distinctes.

Les deux types de textes dégagés pour le discours syndical, très spécifiques, ne s'intègrent pas immédiatement dans ceux proposés par Biber, qui sont pourtant conçus pour rendre compte d'une grande diversité d'énoncés. La question de la généralité des typologies induites à partir des comportements observées reste donc encore largement ouverte.

[24] La présentation de la technique probabiliste correspondante est effectuée dans le chapitre IX.

[25] Soulignons deux apports de ce programme. La simple lecture ne perçoit qu'une partie limitée des sur-emplois effectifs. Elle est bien en peine de juger s'ils sont significatifs ou non. Les sous-emplois, le « creux » d'une partie au regard de l'ensemble, échappent le plus souvent à la conscience. Ils sont ici dégagés.

[26] La CFTC, centrale chrétienne, s'est transformée en 1964 en CFDT, une minorité constituant la CFTC « maintenue ».

[27] L'évolution récente de la CFDT vers plus de pragmatisme s'accompagne d'ailleurs d'une utilisation en congrès de formes proches de celles de la résolution circonstancielle.

3. ÉTIQUETAGE INTEGRAL ET SOCIO-STYLISTIQUE

Un étiquetage intégral bien que rudimentaire permet d'examiner les « parlures » d'un corpus regroupant des énoncés de plusieurs locuteurs de différentes catégories sociales.

3.1 Repérer les catégories et les suites de catégories de différents locuteurs

Enfants, une fois étiqueté et lemmatisé, a été étudié (Habert et Salem, 1995) sous l'angle de l'opposition entre locuteurs sans diplôme, titulaires du baccalauréat et personnes ayant suivi des études supérieures. Un des objectifs de l'utilisation d'une version étiquetée du corpus était de dégager des éléments caractéristiques des « parlures », des styles sociaux présents. Quelles sont les catégories morpho-syntaxiques privilégiées par chaque type de locuteur ? Quels sont les « patrons syntaxiques » qui leur sont propres ?

3.2 Varier le jeu d'étiquettes selon les phénomènes observés

Le corpus a été étiqueté par l'étiqueteur AlethCat[28]. Le « nettoyage manuel » qui a suivi a permis de rectifier un certain nombre d'erreurs de catégorisation et d'homogénéiser la lemmatisation des formes[29].

Les étiquettes employées par l'étiqueteur utilisé, une soixantaine au total, sont relativement rudimentaires : partie du discours, éventuellement sous-type dans la partie du discours (type de déterminant par exemple), traits morphologiques (verbe conjugué / infinitif / participe ..., genre, nombre, personne ...)[30]. Bien d'autres informations pourraient être associées aux « mots » : type de verbe (auxiliaire, modal ...), mot attendant des arguments...

A l'inverse, la présence de certaines indications (le genre, le nombre pour les noms et les adjectifs par exemple) peut rendre plus difficile la perception de certaines régularités : on disperse par exemple les occurrences de la catégorie des noms en masculin singulier, masculin pluriel, féminin singulier et féminin pluriel. Pour faciliter l'étude de telle ou telle opposition, on a donc transformé[31] le jeu d'étiquettes employé, soit en éliminant des informations présentes soit en en rajoutant.

[28] Développé par la société GSI-ERLI. Cet étiqueteur est conçu pour préparer le travail d'un analyseur syntaxique automatique.

[29] Notons que l'étiquetage automatique aboutit parfois à « souder » physiquement des constituants de « mots composés » (*bien_que*, *met_en_évidence*, *vis_à_vis_de* ...).

[30] Une étiquette spécifique non-réponse rend compte de l'absence d'une réponse à la question pour un locuteur donné.

[31] C'est un changement systématique ou à confirmer au coup par coup qu'on pourrait partiellement réaliser avec les fonctions de remplacement d'un simple traitement de textes.

Si l'on prend la phrase suivante :

> je ne sais pas, les gens sont égoïstes peut-être.

en faisant abstraction du lemme, après étiquetage et correction :

> <S01=31> je {PROPERS} ne {ADVNEG} sais {VIPR1S} pas {ADVNEG} , {PONCT-FAIBLE} les {DETDEF} gens {NOMMP} sont {VIPR3P} égoïstes {ADJMP} peut-être {ADV} . {PONCT-FORTE}

que l'on peut représenter aussi, pour plus de « clarté », de la manière suivante :

> <diplôme=études-supérieures, âge=-30>
>
> {forme=je, catégorie=pronom, type=personnel}
>
> {forme=ne, catégorie=adverbe, type=négation}
>
> {forme=sais, catégorie=verbe, mode=indicatif, temps=présent, nombre=singulier, personne=1}
>
> {forme=pas, catégorie=adverbe, type=négation}
>
> [...]

plusieurs transformations ont été utilisées :

- la réduction aux parties du discours traditionnelles :

> {diplôme=études-supérieures, âge=-30}
>
> {forme=je, catégorie=pronom}
>
> {forme=ne, catégorie=adverbe}
>
> {forme=sais, catégorie=verbe}
>
> {forme=pas, catégorie=adverbe}
>
> [...]

- l'élimination des marques de personne, genre et nombre pour les noms et les adjectifs :

> {diplôme=études-supérieures, âge=-30}
>
> [...]
>
> {forme=les, catégorie=déterminant, type=défini}
>
> {forme=gens, catégorie=nom}
>
> {forme=sont, catégorie=verbe, mode=indicatif, temps=présent}
>
> {forme=égoïstes, catégorie=adjectif}
>
> [...]

- l'ajout de la distinction entre adjectifs qualificatifs et adjectifs relationnels :
  Certains adjectifs sont en étroite correspondance avec des noms. Leur

étude complète donc celle de la répartition de cette catégorie majeure au sein du corpus. Ce sont les adjectifs relationnels. Rappelons leurs propriétés (Melis-Puchulu, 1991). Ce sont des adjectifs dénominaux : ils peuvent être mis en rapport avec des séquences *de* + nom comme dans *élection présidentielle* / *élection du président*. Ils ne sont pas gradables : **une carte très géographique*, et ne peuvent être employées de manière prédicative *: *cette carte est géographique*. Dans une séquence d'adjectifs post-posés, ils sont immédiatement après le nom, les adjectifs qualificatifs venant après *: une élection présidentielle surprenante* / **une élection surprenante présidentielle*. L'opposition n'est pas une opposition de nature, mais d'emploi. Ainsi, certains adjectifs relationnels ont également des emplois qualificatifs[32] *: *Cette politique est économique* / *Cette formule est très économique*.

Le résultat est ici :

{diplôme=études-supérieures, âge=-30}

[...]

{forme=les, catégorie=déterminant, type=défini}

{forme=gens, catégorie=nom}

{forme=sont, catégorie=verbe, mode=indicatif, temps=présent, nombre=pluriel, personne=3}

{forme=égoïstes, catégorie=adjectif, type=qualificatif}

[...]

Ces transformations, une fois effectuées, ont été soumises à l'analyse quantitative les différentes versions étiquetées du texte réduites à leurs seules étiquettes, ce qui donne pour l'étiquetage en parties du discours :

{diplôme=études-supérieures, âge=-30}

{catégorie=pronom}

{catégorie=adverbe}

{catégorie=verbe}

{catégorie=adverbe}

{catégorie=ponctuation}

[...]

ou encore en éliminant le nom du trait retenu :

{diplôme=études-supérieures, âge=-30}

{pronom}

{adverbe}

{verbe}

[32] Et inversement, certains adjectifs d'emploi surtout qualificatif peuvent se révéler relationnels selon le contexte. On trouve ainsi dans ***Menelas*** *syndrome douloureux thoracique*, où la place de *douloureux* entre le nom et un autre adjectif relationnel prouve que cet adjectif est ici relationnel. **Syndrome très douloureux* est d'ailleurs impossible dans ce domaine.

{adverbe}

{ponctuation}

[...]

Le programme d'analyse des sur-emplois et des sous-emplois évoqué *supra* permet d'opposer les locuteurs selon leur niveau d'études. Ce sont les étiquettes, pour chacun des jeux, qui sont soumises à examen, mais aussi les suites d'étiquettes, les segments répétés[33] constitués d'étiquettes. Une fois dégagées les tendances d'emploi des étiquettes et de leurs enchaînements, des outils de filtrage permettent d'extraire dans les textes catégorisés les séquences relevant des schémas syntaxiques retenus.

3.3 Une première opposition : style nominal et style verbal

L'examen des proportions relatives d'emploi des parties du discours selon les parties du corpus est instructive. La proportion des noms et des adjectifs croît avec le niveau de diplôme. À l'inverse, le domaine du verbal (verbes, adverbes, pronoms) décroît avec l'élévation du niveau d'études. Ce constat rejoint d'ailleurs ceux faits sur plusieurs corpus pour d'autres études socio-linguistiques. On notera la place éminente, toutes parties confondues, du nom (et des prépositions) : elle tient peut-être à ce que le type de question posée favorise des énoncés qui se présentent sous la forme d'un groupe nominal.

Si l'on s'en tient aux parties du discours seules et qu'on exclut les segments répétés dans lesquels elles entrent, les sans-diplômes se caractérisent par les non-réponses et par le sur-emploi du verbe (et des catégories associées : adverbe et pronom), les plus diplômés par le suremploi des adjectifs et de la coordination. Le faible nombre des catégories employées et le nombre important d'occurrences de chaque étiquette débouchent sur des segments répétés d'étiquettes extrêmement nombreux. On note la présence de syntagmes prépositionnels enchaînés chez les bacheliers comme : [{nom} {préposition} {déterminant} {nom} {préposition} {déterminant} {nom}], ainsi que le poids des adjectifs chez les diplômés du supérieur, en particulier dans des coordinations :

[{nom} {adjectif} {coordonnant} {adjectif}]

[{adjectif} {ponctuation}[34] {adjectif}]

[{nom} {adjectif} {ponctuation} {nom} {adjectif}]

[{déterminant} {nom} {adjectif} {ponctuation} {déterminant} {nom} {adjectif}]

La réduction du corpus aux seules parties du discours fournit une première approche de l'utilisation du matériel linguistique selon les types

[33] L'utilisation de segments répétés de formes ou d'étiquettes est présentée dans le chapitre IX.

[34] Ici comme dans les deux segments répétés suivants, il s'agit en fait de la virgule, dans son rôle de coordonnant.

de locuteurs. Certains phénomènes se trouvent cependant « écrasés » par cette réduction : le sur-emploi significatif de la catégorie adverbe chez les non-diplômés correspond dans près de la moitié des cas (354 occurrences sur 653) à des adverbes de négation. C'est sont alors les résultats obtenus avec un jeu d'étiquettes à mi-chemin du jeu restreint des parties du discours et de celui, trop éclaté, fourni par l'étiqueteur AlethCat qui ont été examinés. Il a semblé important de pouvoir disposer de sous-types des catégories « majeures » employées (à l'instar d'adverbe de négation par rapport à adverbe).

3.4 Examen des patrons syntaxiques caractéristiques de chaque type de locuteur

Cumulant des emplois multiples, les unités lexicales de la négation (*ne*, *pas*, *guère*, *jamais*, *que*) dominent les énoncés des sans-diplômes. Elles structurent le patron sur-employé [{pronom personnel} {adverbe négation} {verbe 1ère personne singulier} {adverbe négation}][35], de type *je ne vois pas, je ne sais pas,* ou le même patron suivi de la virgule : *je ne sais pas,* <reste de la réponse>, qui ne constitue pas exactement une réponse, mais une dévalorisation préalable de la réponse à venir. Par ailleurs, bon nombre d'exemples du patron [{adverbe négation} {verbe présent 3ème personne singulier} {adverbe négation}], sur-employé également par ces locuteurs, correspondent à l'indication par l'enquêteur de la difficulté à répondre chez la personne interrogée : *ne voit pas de raisons* (4 occurrences), *ne sait pas* (8 occurrences) et des variantes comme *ne peut pas répondre*. La non-réponse, comme silence (non-réponse), comme refus explicite de répondre, ou comme mise en doute préalable[36] des propos tenus, est centrale dans cette partie. Le patron sur-employé [{pronom personnel} {adverbe négation } {verbe 3ème personne singulier} {adverbe négation}] est dû pour l'essentiel à l'emploi négatif du présentatif *il y a* dans des réponses comme : *c'est la situation qui décide le logement si il n'y a pas de place* ou encore *quand il n'y a pas assez d'argent dans le ménage*. Il ne faut pas cependant en tirer des conséquences quant à l'orientation argumentative des réponses. Même si l'on trouve des séquences qui mentionnent des difficultés (*il n'y a pas de travail*, 3 occurrences dont l'une en contexte conditionnel), le présentatif négatif peut servir au locuteur à plaider au contraire pour le fait d'avoir des enfants. Les réponses suivantes en témoignent*: il n'y a pas de couple sans enfant* ou encore *il n'y a pas de raison valable*. Les sans- diplômes se caractérisent en outre par des phrases plus courtes, éventuellement réduites à un nom seul, non déterminé[37].

[35] Nous ne donnons que les valeurs des traits pour faciliter la lecture.
[36] Pour déterminer la place du phénomène, au début de chaque réponse a été introduit un « anti-point », noté {ponctuation début-phrase}. Les segments répétés comprenant cette étiquette confirment la tendance des locuteurs sans diplôme à commencer la phrase par un pronom personnel (en règle générale la première personne du singulier) suivi d'une négation : [{ponctuation début-phrase} {pronom personnel} {verbe indicatif présent 1ère personne singulier} {adverbe négation}{ponctuation faible}].
[37] Le motif [{nom} {ponctuation forte}] est sur-employé, les formes correspondantes les plus employées étant *chômage*, 11 occurrences, *égoïsme*, 8 occurrences et *argent*, 7 occurrences.

Les bacheliers sont caractérisés par les enchaînements de syntagmes prépositionnels, puisqu'on trouve des patrons comme : [{nom} {préposition} {article défini} {nom} {préposition} {article défini} {nom}] ou encore comme [{nom} {adjectif} {ponctuation faible} {nom} {préposition} {nom}]. Ce dernier patron est lié à des énumérations nominales, non déterminées (cf. l'absence de déterminant après la ponctuation faible) comme dans la réponse *: raison financière, situation de travail, peur de perdre son travail pour la femme qui s'absente pour raison de maternité*.

Les plus diplômés privilégient nettement l'adjectif et une forme qui en est proche, le participe passé, en particulier dans des coordinations, dans des patrons répétés comme [{nom} {adjectif} {coordonnant} {adjectif}] ou [{adjectif} {ponctuation faible} {adjectif}].

3.5 Préciser l'emploi des adjectifs : qualificatifs et relationnels

Hors contexte, les lemmes des adjectifs du corpus ont été répartis entre les catégories suivantes : {adjectif qualificatif} (*mauvais*), {adjectif relationnel} (*géographique*) et {adjectif qualificatif/relationnel} (*économique*). Cet enrichissement des étiquettes des adjectifs a ensuite été appliqué au texte : l'étiquette {adjectif} associée à *égoïstes* devient par exemple {adjectif qualificatif}. L'examen de la répartition des adjectifs relationnels par rapport aux qualificatifs permet de préciser le fonctionnement dans le corpus de la catégorie nominale prise au sens large .

Les adjectifs n'apparaissent pas dans les formes et segments sur-employés des non-diplômés. En ce qui concerne les bacheliers, seule la catégorie {adjectif relationnel} apparaît comme sur-employée, isolée ou dans des segments répétés. Ce sur-emploi souligne la nature nominale et prépositionnelle de cette partie (puisqu'un adjectif relationnel est équivalent à un syntagme prépositionnel). Cette équivalence est particulièrement flagrante dans le segment répété [{nom} {adjectif relationnel} {ponctuation faible} {nom} {préposition} {nom}] qui coordonne (par une virgule) un nom modifié par un adjectif relationnel et un nom dominant un syntagme prépositionnel. L'adjectif relationnel caractérise davantage les diplômés du supérieur. L'examen des contextes montre en effet que les adjectifs portant l'étiquette {adjectif qualificatif/relationnel} sont en fait tous relationnels dans cette partie : les constats quantitatifs sous-estiment donc la place des adjectifs relationnels. Voici quelques segments répétés significatifs :

[{déterminant défini} {nom} {adjectif relationnel}]

[{nom} {adjectif qualificatif/relationnel}]

3.6 Evaluation et perspectives

L'analyse des décomptes portant sur l'utilisation de divers jeux d'étiquettes donne une image intéressante de l'usage de l'appareil

linguistique par les différents ensembles de locuteurs : expression personnelle, modalisant la réponse faite, à dominante négative pour les sans-diplômes *versus* expression nominale, située hors du *ici et maintenant* pour les diplômés. Les bacheliers marquent une préférence pour les syntagmes prépositionnels, les diplômés du supérieur pour les adjectifs en particulier coordonnés. Les locuteurs ayant fait des études supérieures font appel plutôt aux adjectifs dénominaux qu'aux syntagmes prépositionnels pour modifier les noms, à l'inverse des locuteurs ne possédant que le baccalauréat. S'agirait-il d'un phénomène d'hypercorrection, d'une manière d'éviter le « style substantif » ?

Cependant, bien d'autres interprétations pourraient être produites pour les données constituées avec ces différents jeux d'étiquettes. Par exemple entre des réponses directes (baccalauréat et études supérieures) et des réponses différées (sans-diplômes), où les formules comme *je ne sais pas,* etc., ressemblent aux items de retardement de la réponse mis en évidence en analyse de la conversation.

4. UTILISER ETIQUETEURS ET CORPUS ETIQUETES

4.1 Adapter l'étiquetage aux objectifs de recherche

4.1.1 Un étiquetage est orienté par une famille de tâches

Meyer et Tenney parlent (1993, p. 25-26) d'étiquetage *finalisé (problem-oriented tagging*), à propos de l'étude de l'apposition dans *Survey of English Usage* faite par l'un d'eux. Ils ajoutent que les programmes d'étiquetage disponibles « sont moins utiles pour le linguiste travaillant sur corpus qui souhaite étudier une construction linguistique donnée en détail et adapter le jeu d'étiquettes qu'il met en œuvre pour étudier cette construction. »

Il faut généraliser ce constat. Un étiquetage est toujours orienté par une tâche, même si c'est implicite. Le jeu d'étiquettes utilisé permet d'étudier certains phénomènes ou de développer certains traitements ultérieurs, tandis qu'il laisse d'autres aspects linguistiques dans l'ombre et n'est pas compatible avec d'autres applications. Ainsi, la distinction du genre et du nombre pour les noms et adjectifs dans l'étiquetage d'***Enfants*** n'est pas forcément pertinente pour une étude énonciative de ce corpus, mais par contre, elle est utile pour une analyse syntaxique ultérieure : elle permet de vérifier des contraintes d'accord au sein du groupe nominal. À l'inverse, tous les étiqueteurs ne fournissent pas le temps et la personne pour les verbes conjugués[38], bien que cette information soit

[38] Ne serait-ce qu'en raison de la difficulté d'assurer une désambiguïsation efficace sur ce point : *travaille* est un présent de l'indicatif, 1ère et 3ème personne du singulier, mais aussi un présent du subjonctif aux mêmes personnes et enfin un impératif 2ème

particulièrement précieuse dans une perspective typologique comme celle de Biber. Une catégorisation donne ainsi à voir certains phénomènes et en ignore d'autres. Il faut donc multiplier les points de vue et à tout le moins être conscient des capacités heuristiques et des angles morts des jeux d'étiquettes auxquels on a recours. Les projets de comparaison et d'évaluation d'étiqueteurs se développent aujourd'hui (Paroubek *et al.*, 1997). Ce qu'on peut en attendre, ce n'est certainement pas une mise en évidence de la « meilleure catégorisation », ce qui n'a pas grand sens, mais l'identification des objectifs, points forts et faiblesses de chaque catégorisation et de l'adéquation de chacune aux projets de recherche envisagés.

4.1.2 Un étiquetage peut être « détourné »

Nous rencontrons avec les corpus étiquetés une situation courante pour les corpus annotés en général. L'annotation du corpus utilisé ne correspond pas exactement à la classification souhaitée des données, aux phénomènes que l'on souhaite isoler, au regard théorique que l'on porte sur eux. Pire : pour diverses raisons (le plus souvent le manque de moyens financiers et humains), il n'est pas possible de ré-étiqueter le corpus.

Il s'agit alors de « composer » avec l'état présent de l'étiquetage, d'en tirer les informations qui se rapprochent de celles recherchées. C'est cette démarche que nous avons vue à l'œuvre dans les études typologiques sur le discours syndical : faute de disposer de corpus étiquetés (il y a 15 ans, dans les limbes pour l'anglais et inexistants pour le français), on étudie aussi précisément que possible un ensemble délimité de formes graphiques (de « mots »), malgré le " bruit " introduit par l'utilisation de cette représentation sommaire.

A l'inverse, si, dans le cas présent, une telle démarche typologique peut se satisfaire, pour un premier dégrossissage, de corpus « bruts », c'est-à-dire réduits à des formes graphiques, elle gagne sans conteste à utiliser des corpus étiquetés de manière spécifique. L'écart entre les données utilisées par ces différentes analyses et la plus ou moins grande immédiateté d'interprétation qui en résulte débouche néanmoins sur la nécessité plus générale de vérifier l'adéquation possible (au prix de détournements éventuels) entre les conventions d'annotation du corpus utilisé et les objectifs de recherche visés.

4.1.3 Le ré-étiquetage est incontournable

L'écart entre les catégories associées à un corpus déjà catégorisé ou fournies par un étiqueteur accessible et celles dont on peut avoir besoin pour une étude donnée implique souvent une recatégorisation (partielle) du corpus. Nous avons montré comment, pour ***Enfants***, l'ajout d'une

---

personne du singulier.

nouvelle distinction (adjectif qualificatif / relationnel) venait préciser une étiquette existante. Le ré-étiquetage peut aussi conduire à des révisions plus drastiques, lorsque les choix de segmentation de départ sont remis en cause (le choix des « mots composés » pertinents pour le corpus en cause) ou quand certains phénomènes sont traités différemment (par exemple, *rapide* analysé tantôt comme un adjectif tantôt comme un adverbe dans *Prenons une rapide décision*).

Le ré-étiquetage total ou partiel peut aussi avoir comme visée l'alignement des résultats de deux étiqueteurs sur un même corpus, à des fins de comparaison ou d'évaluation (Atwell *et al.*, 1994). Selon Belmore (1994, p. 52) : « Une manière d'utiliser les corpus pour améliorer de manière cumulative les analyses consiste à déterminer les différences exactes entre deux analyses d'un même corpus. Dans l'idéal, l'une des deux analyses partirait de la première et représenterait alors un essai explicite d'amélioration. »

4.2 Environnements de catégorisation et de manipulation de texte étiqueté

Paradoxalement, il semble que le besoin d'environnements informatiques de catégorisation et de manipulation de texte étiqueté, souvent souligné par les participants des projets d'étiquetage et de structuration de corpus, reçoive dans l'immédiat peu de réalisations concrètes (Greenbaum et Yibin, 1994, p. 44).

4.2.1 Catégoriser

On peut vouloir étiqueter, totalement ou partiellement, un texte « nu ». S'il s'agit d'utiliser un corpus déjà étiqueté ou les résultats d'un étiqueteur disponible, la finesse des distinguos nécessaires pour des analyses proprement linguistiques suppose des programmes permettant de préciser l'étiquetage morpho-syntaxique accompagnant désormais nombre de corpus. Elle implique aussi des modules de catégorisation interactive ou de modification interactive d'étiquetages préalables, certaines valeurs d'étiquettes ne pouvant pas être attribuées automatiquement[39].

4.2.2 Manipuler des corpus étiquetés

Les programmes nécessaires ici permettent d'extraire du texte étiqueté des motifs arbitrairement complexes. Les constituants de ces motifs sont,

[39] Par exemple, la distinction entre déterminants définis spécifiques *versus* génériques dans (Sueur, 1982).

ici encore, des structures de traits[40]. Le motif (ou patron) correspondra au fragment de texte pour lequel les structures de traits de ses composants s'apparient avec celles des éléments correspondants du texte. On parle de filtrage (*pattern-matching*). Des opérateurs permettent la conjonction, la disjonction, l'optionalité, la répétition de ces contraintes, etc. Par exemple, le motif :

> [{nom} {adjectif relationnel ∨ qualificatif/relationnel} {coordonnant} {adjectif relationnel ∨ qualificatif/relationnel}]

permet de chercher les noms suivis de deux adjectifs coordonnés soit relationnels soit qualificatifs ou relationnels (c'est ce qu'indique la disjonction ∨).

De tels environnements facilitent le nécessaire retour au contexte qui permet d'éviter les commentaires oiseux de simples artefacts. Dans ***Enfants***, par exemple, les réponses fournies par les plus diplômés paraissent plus riches en séquences du type : [{adjectif qualificatif} {nom}]. Ce résultat attire l'attention : en français moderne, l'antéposition de l'adjectif est une construction de langue tenue. Déception : l'examen des séquences relevant de ce patron montre qu'en fait, il s'agit souvent d'adjectifs modifiant un nom antéposé. L'ambiguïté est due à l'absence de marque de ponctuation entre les groupes nominaux dans des suites de formes comme : *temps libre argent*.

5. ENJEUX THEORIQUES

5.1 Le dit est le dire

L'examen des catégories employées, et des segments de catégories conduit à s'attacher aux patrons syntaxiques des énoncés, voire aux genres textuels qui peuvent expliquer le recours à tel type de construction. D'autres phénomènes linguistiques s'offrent à une exploration méthodique et à la quantification. Nous espérons avoir montré qu'une analyse du dire (du style, du mode de parler) était tout aussi instructive qu'une analyse du dit. Le détour par des catégories abstraites, ici morpho-syntaxiques, introduit une bienfaisante étrangeté dans l'appréhension du corpus. Ce pas de côté contrebalance la trompeuse immédiateté des formes lexicales, dont le sens s'impose trop évidemment. Mais, en même temps, certaines associations de traits dans les dimensions dégagées par Biber ou le sur-emploi de telle étiquette dans l'étude d'***Enfants*** demeurent énigmatiques. On ne dispose pas forcément dans l'immédiat des cadres théoriques nécessaires pour examiner les données ainsi produites.

[40] Là encore, comme en 1.3, nous fournissons une représentation unifiée des différentes possibilités effectives dans tel ou tel système d'interrogation de texte étiqueté.

5.2 Linguistique et textualité

Benveniste assignait à la linguistique la phrase comme horizon d'analyse. Il n'en a pas moins exploré les régularités proprement textuelles liées à l'utilisation de l'appareil de l'énonciation. Sa distinction histoire / discours a donné naissance à des typologies ou des grilles d'analyse plus fines. En didactique du français, cette dichotomie a été mobilisée largement pour aider les apprenants à maîtriser les conditions de bonne formation des textes.

L'utilisation de corpus étiquetés diversifiés offre désormais la possibilité d'examiner sérieusement l'hypothèse que les textes effectifs relèvent de types fondamentaux qui expliquent un certain nombre de leurs traits linguistiques. Les études existantes en fournissent des caractérisations empiriques fines. La généralité des catégories dégagées, leur lien aux genres et registres intuitivement distingués par les locuteurs restent à travailler. Ces résultats appellent peut-être un renouveau de la linguistique textuelle : on attend un modèle de la compétence textuelle qui intègre les contraintes détaillées mises en évidence.

Il reste également à explorer le fonctionnement social des types de textes disponibles dans une communauté langagière donnée. Pour Bakhtine : « Nous ne parlons qu'à travers certains genres discursifs, c'est-à-dire que tous nos énoncés possèdent certaines formes relativement stables et typiques pour se constituer en totalités » (Todorov, 1981, p. 129). L'organisation de chacun de ces genres est socialement significative : « Le genre forme [...] un système modélisant qui propose un simulacre du monde » (*ibid.,* p. 128). C'est le cas des deux types de résolutions de congrès syndicaux évoqués en 2.3. La résolution déclarative (ou circonstancielle) va de pair avec un refus de tenir un discours global sur la société. Elle légifère essentiellement pour le laps de temps qui la sépare du congrès suivant. La résolution analytique (ou théorique) s'installe dans l'éternel présent de la théorie, dépassant les limites du *ici et maintenant*. L'idéologie s'y exprime dans de longs développements sans locuteur explicite. Les articulations logiques veulent entraîner dans un réseau d'enchaînements qui font appel au simple examen de la « nature des choses »[41].

5.3 Analyses multi-dimensionnelles

Toutes proportions gardées, les études typologiques à partir de formes graphiques seules ou à partir de traits linguistiques clairement identifiés (cf. 2) tiennent de la reconstitution d'animaux disparus à partir de fossiles épars et incomplets. Au vu de données langagières fragmentaires, on

[41] Il est intéressant à cet égard de noter que l'utilisation par la CFTC et la CFDT de ces deux types de résolutions ne se superpose pas mécaniquement à l'évolution historique de cette confédération. La déconfessionnalisation de 1964 n'entraîne pas un changement sur ce plan. C'est quand cette organisation veut affirmer fortement un projet social propre qu'elle recourt à la résolution théorique : en 1945 — dans l'immédiat après-guerre, et après 1968.

postule l'existence d'un squelette syntaxique[42] voire textuel. On fait l'hypothèse de dépendances fonctionnelles entre des éléments relevant de niveaux distincts de l'analyse linguistique. Avec le risque d'inventer des « monstres langagiers » sans existence réelle.

Les techniques d'analyses statistiques multi-dimensionnelles comme l'analyse factorielle des correspondances utilisée par Biber ont précisément pour objectif de manifester les corrélations effectives entre des variables multiples. Elles mettent en évidence des régularités qui échappent à l'observation « à l'œil nu ». Elles débouchent sur des regroupements de comportements langagiers qui peuvent renouveler nos analyses des dépendances entre niveaux linguistiques[43]. Elles manifestent des oppositions qui restructurent notre catégorisation préalable des données.

[42] Comme la proto-phrase donnée comme sous-jacente aux résolutions déclaratives (Bergounioux *et al.*, 1982, p. 178) évoquées en 2.3 : « considérant [...] le congrès [...] {verbe déclaratif 3ème personne présent} [...] que [...] {subjonctif} [...] {déterminant indéfini}. »

[43] Pour poursuivre la métaphore, notons que l'apport de ces méthodes a été considérable en classification des espèces : elles ont permis d'améliorer les taxonomies existantes, limitées dans leur capacité à percevoir et organiser des corrélations multiples.